CASC research team earns patent on data mining software

Pete Eltgroth, CASC Director

January 30, 2004

CASC researchers Chandrika Kamath and Erick Cantu-Paz were recently granted a U.S. Patent for a data mining system that uncovers patterns, associations, anomalies, and other statistically significant structures in data. Data mining has been defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data." Mining scientific data is a process that uses techniques from image processing, machine learning, statistics, and visualization to discover knowledge in data and present it in a form which is easily comprehensible to humans.

Kamath and Cantu-Paz filed a patent application on June 8, 2001 for a Parallel Object-oriented Data Mining System. U.S. Patent No. 6,675,164 B2 was issued on January 6, 2004. This work was funded by LDRD and ASCI VIEWS. It is part of the Sapphire project in CASC which conducts research in a variety of techniques to discover new knowledge in diverse data sources. This invention provides a data mining system that uncovers patterns, associations, anomalies, and other statistically significant structures in data.

The idea patented is the system architecture for the Sapphire project. Sapphire was designed to take into account several aspects of data mining in order to achieve good performance on real data. Not all problems require the entire data mining process, so each step in Sapphire is modular and capable of standalone operation. Not all algorithms are suitable for a given problem, so the software includes several algorithms for each task and allows easy plug-and-play of these algorithms. Each algorithm typically depends on several parameters, so the software allows user-friendly access to these parameters. Intermediate data is stored appropriately to support refinement of the data mining process. And, the knowledge-domain dependent and independent parts are clearly identified to allow maximum re-use of software so the system can support several different applications.

The Sapphire architecture supports these design goals through several parallel, object-oriented modules that may be combined as appropriate. The basic system consists of modules for reading, writing, and displaying input data in different formats, identifying objects in the data, extracting features for these objects, several pattern recognition techniques, and storage of the extracted features. This system may be enhanced with modules for data fusion, multi-resolution, de-noising, sampling, and dimension reduction, among others.

See the Sapphire web site for more information.